ReDet: A Rotation-equivariant Detector for Aerial Object Detection

Jiaming Han*, Jian Ding*, Nan Xue, Gui-Song Xia
Wuhan University, Wuhan, China
{hanjiaming, jian.ding, xuenan, guisong.xia}@whu.edu.cn

Abstract

Recently, object detection in aerial images has gained much attention in computer vision. Different from objects in natural images, aerial objects are often distributed with arbitrary orientation. Therefore, the detector requires more parameters to encode the orientation information, which are often highly redundant and inefficient. Moreover, as ordinary CNNs do not explicitly model the orientation variation, large amounts of rotation augmented data is needed to train an accurate object detector. In this paper, we propose a Rotation-equivariant Detector (ReDet) to address these issues, which explicitly encodes rotation equivariance and rotation invariance. More precisely, we incorporate rotation-equivariant networks into the detector to extract rotation-equivariant features, which can accurately predict the orientation and lead to a huge reduction of model size. Based on the rotation-equivariant features, we also present Rotation-invariant RoI Align (RiRoI Align), which adaptively extracts rotation-invariant features from equivariant features according to the orientation of RoI. Extensive experiments on several challenging aerial image datasets DOTA-v1.0, DOTA-v1.5 and HRSC2016, show that our method can achieve state-of-the-art performance on the task of aerial object detection. Compared with previous best results, our ReDet gains 1.2, 3.5 and 2.6 mAP on DOTA-v1.0, DOTA-v1.5 and HRSC2016 respectively while reducing the number of parameters by 60% (313 Mb vs. 121 Mb). The code is available at: https://github.com/csuhan/ReDet.

1. Introduction

This paper studies the problem of object detection in aerial images, a recently-emerged challenging problem in computer vision [35]. Different from objects in nature images, objects in aerial images are often distributed with arbitrary orientation. To cope with these challenges, aerial object detection are usually formulated as an oriented object detection task by relying on Oriented Bounding Boxes (OBBs) representation instead of using Horizontal Bounding Boxes (HBBs) [7, 35, 38, 40].

Recently, many well-designed oriented object detectors have been proposed and reported promising results on challenging aerial image datasets [21, 35]. In order to achieve accurate object detection in unconstrained aerial images, most of them are devoted to extract rotation-invariant features [7, 10, 22, 37]. In practice, Rotated RoI (RRoI) warp-

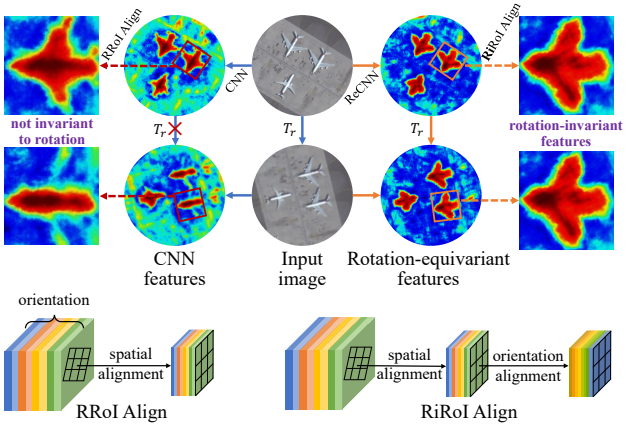

Figure 1. **Illustration of our method (top) and comparisons of RRoI warping (bottom)**. CNN features are not equivariant to the rotation T_r, *i.e.*, feeding a rotated image to CNNs is not the same as rotating feature maps of the original image. Therefore, the corresponding RoI features are not invariant to rotation. In contrast, our method adopts rotation-equivariant CNNs (ReCNN) to extract rotation-equivariant features. Let I and Φ be the input and ReCNN respectively, the equivariance of our method can be expressed as: $\Phi(T_r I) = T_r \Phi(I)$, *i.e.*, applying a rotation T_r to the image I is the same as the rotation of features. Since we have obtained rotation-equivariant features, rotation-invariant features can be extracted by RRoI warping. While RRoI Align can only achieve rotation invariance in the spatial dimension, we present a novel Rotation-invariant RoI (RiRoI) Align to extract rotation-invariant features in both spatial and orientation dimensions.

The study of this paper is funded by the National Natural Science Foundation of China (NSFC) under grant contracts No.61922065, No.61771350 and No.41820104006 and 61871299. It is also supported by Supercomputing Center of Wuhan University.

*Equal contribution.

Corresponding author: Gui-Song Xia (guisong.xia@whu.edu.cn).

ing (*e.g.*, RRoI Pooling [22] and RRoI Align [7]) is the most commonly used method to extract rotation-invariant features, which can warp region features precisely according to the bounding boxes of RRoI in the 2D planar. However, RRoI warping with regular CNN features can not produce exactly rotation-invariant features. The rotation invariance is approximated by employing larger capacity networks and more training samples to model the rotation variation. As shown in Fig. 1, the regular CNNs are not equivariant to the rotation, *i.e.*, feeding a rotated image to CNNs is not the same as rotating feature maps of the original image. Therefore, region features warped from regular CNN feature maps are usually unstable and delicate as the orientation changes.

Some recently proposed methods [5, 13, 33] extend CNNs to larger groups and achieve rotation equivariance[1] with group convolutions [5]. Feature maps of these methods have additional orientation channels recording features from different orientations. However, directly applying the ordinary RRoI warping to rotation-equivariant features is unable to produce rotation-invariant features, as it can only warp region features in the 2D planar, *i.e.*, the spatial dimension, while the orientation channels are still misaligned. To extract completely rotation-invariant features, we also need to adjust the orientation dimension of feature maps according to the orientation of RRoI.

In this paper, we propose a Rotation-equivariant Detector (ReDet) to extract completely rotation-invariant features from rotation-equivariant features. As shown in Fig. 1, our method consists of two parts: rotation-equivariant feature extraction and rotation-invariant feature extraction. Firstly, we incorporate rotation-equivariant networks into the backbone to produce rotation-equivariant features, which can accurately predict the orientation and reduce the complexity of modeling orientation variations. Since directly apply the RRoI warping still cannot extract rotation-invariant features from the rotation-equivariant features, we propose a novel Rotation-invariant RoI Align (RiRoI Align). It can warp region features according to the bounding boxes of RRoI in the spatial dimension and align features in the orientation dimension by circularly switching orientation channels and feature interpolation. Finally, the combination of rotation-equivariant backbone and RiRoI Align forms our ReDet to extract completely rotation-invariant features for accurate aerial object detection.

Extensive experiments performed on the challenging aerial image datasets DOTA [35] and HRSC2016 [21] demonstrate the effectiveness of our method. We summary our contributions as: **(1)** We propose a Rotation-equivariant Detector for high-quality aerial object detection, which encodes both rotation equivariance and rotation invariance. To our best knowledge, it is the first time that rotation equivariance has been systematically introduced into oriented object detection. **(2)** We design a novel RiRoI Align to extract rotation-invariant features from rotation-equivariant features. Different from other RRoI warping methods, RiRoI Align produces completely rotation-invariant features in both spatial and orientation dimensions. **(3)** Our method achieves the state-of-the-art **80.10**, **76.80** and **90.46** mAP on DOTA-v1.0, DOTA-v1.5 and HRSC2016, respectively. Compared with previous best results, our method gains **1.2**, **3.5** and **2.6** mAP improvements. Compared with the baseline, our method shows consistent and substantial improvements and reduces the number of parameters by **60%** (313 Mb *vs.* 121 Mb). Moreover, our method achieves better model size *vs.* accuracy trade-off (shown in Fig. 2).

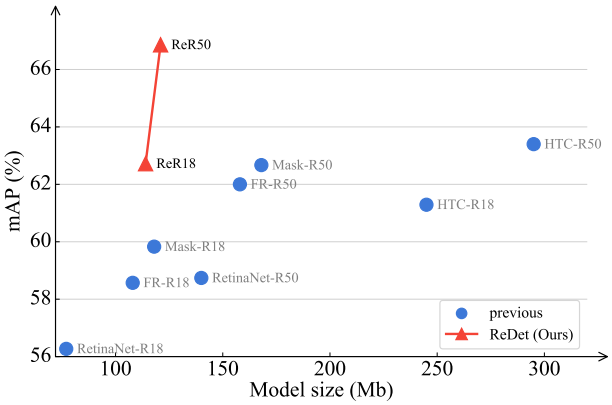

Figure 2. **Model size *vs.* accuracy (mAP) on DOTA-v1.5.** We evaluate RetinaNet OBB [18], Faster R-CNN OBB (FR) [27], Mask R-CNN (Mask) [11] and Hybrid Task Cascade (HTC) [2] with ResNet18 (R18) and ResNet50 (R50) backbones. Note all algorithms are our re-implemented version for DOTA, which is consistent with Tab. 7. Our ReDet is tested with ReResNet18 (ReR18) and ReResNet50 (ReR50) backbones. Compared with other methods with R18/R50 backbones, our ReDet with a ReR18 backbone achieves competitive performance. Using a deeper backbone (ReR50), our ReDet outperforms all methods by a large margin and achieves better model size *vs.* accuracy trade-off.

2. Related Works

2.1. Oriented Object Detection

Unlike most general object detectors [8, 9, 18, 20, 26, 27, 44] that use HBBs, oriented object detectors locate and classify objects with OBBs, which provide more accurate orientation information of objects. This is essential for detecting aerial objects with large aspect ratio, arbitrary orientation and dense distribution. With the development of general object detection, many well-designed methods have been proposed for oriented object detection [1, 7, 24, 35, 38, 40, 42], showing promising performance on challenging datasets [21, 35]. To detect objects with arbitrary orientation, some methods [1, 22, 43] adopt numerous rotated an-

[1]Equivariance is a property that applying transformations to the input produces transformations of the feature in a predictable way.

chors with different angles, scales and aspect ratios for better regression while increasing the computation complexity. Ding *et al.* proposed RoI Transformer [7] to transform Horizontal RoIs (HRoIs) into RRoIs, which avoids a large number of anchors. Gliding vertex [36] and CenterMap [30] use quadrilateral and mask to accurately describe oriented objects, respectively. R^3Det and S^2A-Net align the feature between horizontal receptive fields and rotated anchors. DRN [24] detects oriented objects with dynamic feature selection and refinement. CSL [38] regards angular prediction as a classification task to avoid discontinuous boundaries problem. Recently, some CenterNet [44]-based methods [24, 31, 41] show their advantages in detecting small objects. The above methods are devoted to improving object representations or feature representations. While our method is dedicated to improving the feature representation throughout the network: from the backbone to the detection head. Specifically, our method produces rotation-equivariant features in the backbone, significantly reducing the complexity in modeling orientation variations. In the detection head, the RiRoI Align extracts completely rotation-invariant features for robust object localization.

2.2. Rotation-equivariant Networks

Cohen *et al.* first proposed group convolutions [5] to incorporate 4-fold rotation equivariance into CNNs. HexaConv [13] extends group convolutions to 6-fold rotation equivariance over hexagonal lattices. To achieve rotation equivariance on more orientations, some methods [23, 45] resampling filters by interpolation, while other methods [32, 33, 34] use harmonics as filters to produce equivariant features in the continuous domain. The above methods gradually extend rotation equivariance to larger groups and achieve promising results on the classification task, while our method incorporates rotation-equivariant networks into the object detector, showing significant improvements on the detection task. To our best knowledge, this is the first time that rotation equivariance has been systematically applied to oriented object detection.

2.3. Rotation-invariant Object Detection

The rotation-invariant feature is important for detecting arbitrary oriented objects. However, CNNs show poor performance in modeling rotation variations, which means that more parameters are needed to encode the orientation information. STN [14] and DCN [6] explicitly model the rotation within the network and have been widely applied to oriented object detection [7, 28, 29]. Cheng *et al.* [4] proposed a rotation-invariant layer that imposes an explicit regularization constraint to the objective. Though the above methods can achieve approximated rotation invariance in the image-level, large amounts of training samples and parameters are needed. Besides, object detection requires instance-level rotation-invariant features. Therefore, some methods [7, 22] extend RoI warping [8] to RRoI warping, *e.g.*, RoI Transformer [7] learns to transform HRoIs to RRoIs and warps region features with a rotated position sensitive RoI Align. However, the regular CNNs are not rotation-equivariant. Therefore, even through the RRoI Align, we still cannot extract rotation-invariant features, as shown in Fig. 1. Different from the aforementioned methods, our method proposes Rotation-invariant RoI Align (RiRoI Align) to extract rotation-invariant features from rotation-equivariant features. Specifically, we incorporate rotation-equivariant networks into the backbone to produce rotation-equivariant features, then the RiRoI Align extracts completely rotation-invariant features from rotation-equivariant features in both spatial and orientation dimensions.

3. Preliminaries

Equivariance is a property that applying transformations to the input produces transformations of the feature in a predictable way. Formally, give a transformation group G and a function $\Phi : X \rightarrow Y$, equivariance can be expressed as:

$$\Phi[T_g^X(x)] = T_g^Y[\Phi(x)] \quad \forall (x, g) \in (X, G), \tag{1}$$

where T_g indicates a group action in the corresponding space. Especially when T_g^Y is identical for all T_g^X, equivariance becomes invariance.

In common, CNNs are known to be translation equivariant. Let T_t denotes an action of the translation group $(\mathbb{R}^2, +)$, and apply it to K-dimension feature maps $f : \mathbb{Z}^2 \rightarrow \mathbb{R}^K$, translation equivariance can be expressed as:

$$[[T_t f] * \psi](x) = [T_t [f * \psi]](x), \tag{2}$$

where $\psi : \mathbb{Z}^2 \rightarrow \mathbb{R}^K$ indicates the convolution filter and $*$ is the convolution operation. Recently proposed methods [5, 13, 33] extend CNNs to large groups, achieving both translation and rotation equivariance. Let H denotes a rotation group, *e.g.*, the cyclic group C_N containing discrete rotations by angles multiple of $\frac{2\pi}{N}$. We can define the group G as the semidirect product of the translation group $(\mathbb{R}^2, +)$ and the rotation group H, *i.e.*, $G \cong (\mathbb{R}^2, +) \rtimes H$. By replacing $x \in (\mathbb{R}^2, +)$ with $g \in G$ in Eq. 2, the rotation-equivariant convolution can be defined as:

$$[[T_g f] * \psi](g) = [T_g [f * \psi]](g). \tag{3}$$

Rotation-equivariant Networks. The regular CNNs consists of a series of convolution layers and enjoy the translation weight sharing. Similarly, rotation-equivariant networks are a stack of rotation-equivariant layers with a higher degree of weight sharing, *i.e.*, both translation and rotation. Formally, let $\Phi = \{L_i | i \in \{1, 2, \cdots, M\}\}$ denotes a network with M rotation-equivariant layers under

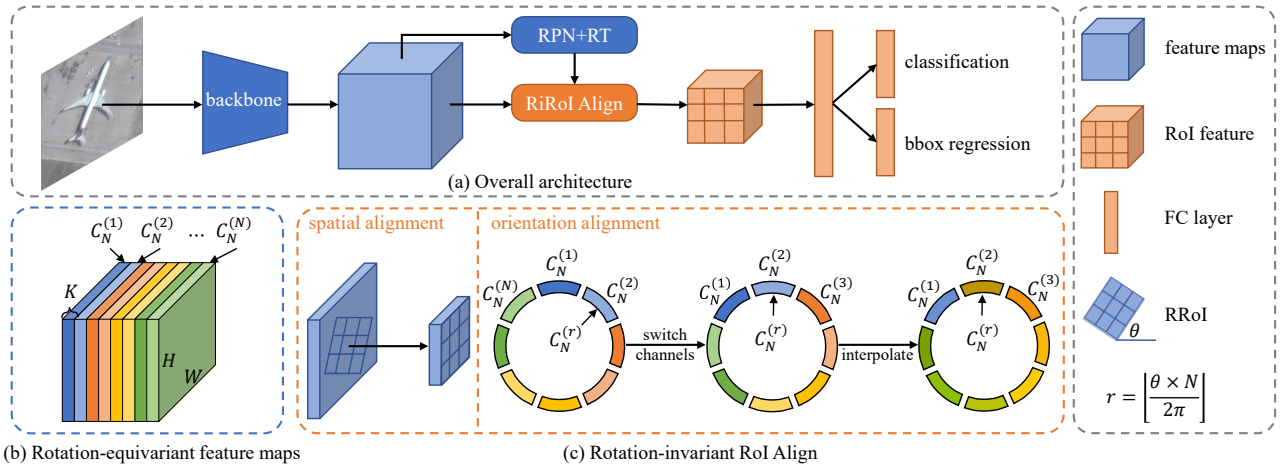

Figure 3. **Overview of our proposed method**. **(a)** Overall architecture of the proposed Rotation-equivariant Detector. We first adopt the rotation-equivariant backbone to extract rotation-equivariant features, followed by an RPN and RoI Transformer (RT) [7] to generate RRoIs. Then we use a novel Rotation-invariant RoI Align (RiRoI Align) to produce rotation-invariant features for RoI-wise classification and bounding box (bbox) regression. **(b)** Rotation-equivariant feature maps. Under the cyclic group C_N, the rotation-equivariant feature maps with the size (K, N, H, W) have N orientation channels, and each orientation channel is corresponding to an element in C_N. **(c)** RiRoI Align. The proposed RiRoI Align consists of two parts: spatial alignment and orientation alignment. For an RRoI (x, y, w, h, θ), spatial alignment warps the RRoI from the spatial dimension, while orientation alignment circularly switches orientation channels and interpolates features to produce completely rotation-invariant features.

group G. For a layer $L_i \in \Phi$, the rotation transformation T_r can be preserved by the layer:

$$L_i[T_r(g)] = T_r[L_i(g)] \quad g \in G. \tag{4}$$

If we apply T_r to the input I and feed it to the network Φ, the transformation T_r[2] will be preserved by the whole network:

$$[\prod_{i=1}^{M} L_i](T_r I) = T_r[\prod_{i=1}^{M} L_i](I). \tag{5}$$

Rotation-invariant Features. For any rotation transformations T_r applied to the input, if its output remains unchanged, we say the output feature is rotation-invariant. Rotation-invariant features can be divided into three levels: image-level, instance-level, and pixel-level. Here we mainly focus on the *instance-level rotation-invariant feature*, which is more suitable for the object detection task. Let $I_R \in I$ and $f_R \in f$ denotes an RoI of the image I and feature maps f ($f = \Phi(I)$), respectively. Assume I_R is a HRoI (x, y, w, h) invariant to the orientation, where (x, y), w and h denote the center point, width and height of the HRoI, respectively. While $T_r I_R$ is an RRoI (x, y, w, h, θ) related to the orientation θ. Similar to Eq. 5, for RoI I_R, the rotation equivariance can be expressed as:

$$\Phi(T_r I_R) = T_r \Phi(I_R). \tag{6}$$

[2]The transformation T_r may have different formulations in different spaces, *e.g.*, the input (image) space and the feature space. Here we do not distinguish it for simplicity. For a deeper discussion of rotation-equivariant networks, we refer the readers to [5] and [33].

If we regard HRoI I_R as the rotation-invariant representation of RRoI $T_r I_R$ in the image I, $\Phi(I_R)$ can be regarded as the rotation-invariant representation of $\Phi(T_r I_R)$ in the corresponding feature space. To get $\Phi(I_R)$, we need to know the rotation transformation T_r. Fortunately, T_r is usually a function of the orientation θ: $T_r = T(\theta)$. In practice, we can simply adopt a RRPN [22] or R-CNN to learn the orientation θ of an RRoI, as well as the transformation T_r. Finally, the rotation-invariant feature $\Phi(I_R)$ can be obtained by applying an inverse transformation $T_r^{'}$ to Eq. 6:

$$\Phi(I_R) = T_r^{'} \Phi(T_r I_R). \tag{7}$$

4. Rotation-equivariant Detector

This section presents details of the proposed Rotation-equivariant Detector (ReDet) to encode both rotation equivariance and rotation invariance. First, we adopt rotation-equivariant networks as the backbone to extract rotation-equivariant features. As discussed before, directly applying the RRoI Align to rotation-equivariant feature maps cannot obtain the rotation-invariant features. Therefore, we design a novel Rotation-invariant RoI Align (RiRoI Align), which produces RoI-wise rotation-invariant features from rotation-equivariant feature maps. The overall architecture of ReDet is shown in Fig. 3. For an input image, we feed it to the rotation-equivariant backbone. Then we adopt RPN to generate HRoIs, followed by an RoI Transformer (RT) [7] that transforms HRoIs to RRoIs. Finally, the RiRoI Align is adopted to extract rotation-invariant features for RoI-wise classification and bounding box regression.

4.1. Rotation-equivariant Backbone

Modern object detectors usually adopt deep CNNs as the backbone to automatically extract deep features with enriched semantic information, *e.g.*, the widely used ResNet [12] with Feature Pyramid Network (FPN) [17]. We also adopt ResNet with FPN as the baseline and implement a rotation-equivariant backbone, named Rotation-equivariant ResNet (ReResNet) with ReFPN.

Specifically, we re-implement all layers of the backbone with rotation-equivariant networks based on `e2cnn` [32], including convolution, pooling, normalization, non linearities, *etc*. Considering the computational budget, ReResNet and ReFPN are only equivariant to the discrete group $(\mathbb{R}^2, +) \rtimes C_N$, *i.e.*, all translations and N discrete rotations. As is shown in Fig. 3 (b), we can feed an image to the rotation-equivariant backbone to produce rotation-equivariant feature maps. Unlike ordinary feature maps, the rotation-equivariant feature maps f with the size (K, N, H, W) have N orientation channels: $f = \{f^{(i)} | i \in \{1, 2, \cdots, N\}\}$, and feature maps of each orientation channel $f^{(i)}$ is corresponding to an element in C_N.

Compared with ordinary backbones, the rotation-equivariant backbone has the following advantages: (a) **Higher degree of weight sharing.** As we have introduced that rotation-equivariant feature maps have an additional orientation dimension. Features from different orientations usually share the same filters with different rotation transformations, *i.e.*, the rotation weight sharing. (b) **Enriched orientation information.** For an input image with a fixed orientation, the rotation-equivariant backbone can produce features from multiple orientations. This is important for oriented object detection, which requires accurate orientation information. (c) **Smaller model size.** Compared with the baseline, we have two choices when designing the backbone: similar computation or similar parameters. Typically, we keep similar computation with the baseline, *i.e.*, preserving the same output channels. Due to the rotation weight sharing, our rotation-equivariant backbone shows a huge reduction of model size, about $1/N$ of parameters.

4.2. Rotation-invariant RoI Align

As introduced in Sec. 3, for an RRoI (x, y, w, h, θ), we can extract rotation-invariant RoI features from rotation-equivariant feature maps with RRoI warping. However, the ordinary RRoI warping can only align features in the spatial dimension, while the orientation dimension leaves misaligned. Therefore, we propose RiRoI Align to extract completely rotation-invariant features. As is shown in Fig. 3 (c), RiRoI Align includes two parts: (a) **Spatial alignment.** For an RRoI (x, y, w, h, θ), spatial alignment warps it from feature maps f to produce rotation-invariant region features f_R in the spatial dimension, which is consistent with RRoI Align [7]. (b) **Orientation alignment**. To ensure RRoIs with different orientations produce completely rotation-invariant features, we perform orientation alignment in the orientation dimension. Specifically, for the output region features $\hat{f}_R$, we formulate orientation alignment as:

$$\hat{f}_R = Int(SC(f_R, r), \theta),\ r = \lfloor \theta N / 2\pi \rfloor, \tag{8}$$

where SC and Int denote the *switching channels* and *feature interpolation* operations, respectively. For the region features f_R, we first calculate an index r, and circularly switch the orientation channels to make sure $C_N^{(r)}$ is the first orientation channel. However, since the rotation equivariance is only achieved in the discrete group C_N, we also need to interpolate the feature if $\theta \notin C_N$. More precisely, we interpolate the orientation feature with its nearest l orientation channels. For example, the output feature of i-th orientation channel with $l = 2$ can be expressed as:

$$\hat{f}_R^{(i)} = (1 - \alpha) f_R^{(i)} + \alpha f_R^{(i+1)}, \tag{9}$$

where $\alpha = \theta N / 2\pi - r$ indicates the distance factor for 1D-interpolation. Note that we use the `mod` function to ensure $i \in [1, N]$ (as well as $i + 1$).

Comparison with RRoI Align+MaxPool. Different from RiRoI Align, warping RoI features with RRoI Align and then maxpooling over the orientation dimension (*i.e.*, orientation pooling) is another approach to extract rotation-invariant features. The orientation pooling operation is usually adopted in classification tasks [5, 33, 45]. For each location in the feature map, it only preserves the orientation with the strongest response, while features from other orientations are abandoned. However, we argue that the response from all orientations, no matter strong or weak, is indispensable for object recognition. In our RiRoI Align, features from all orientations are preserved and aligned with the orientation alignment operation. We will conduct experiments to show the advantage of our RiRoI Align in Sec. 5.

5. Experiments and Analysis

5.1. Datasets

DOTA [35] is the largest dataset for oriented object detection in aerial images with two released versions: **DOTA-v1.0** and **DOTA-v1.5**. **DOTA-v1.0** contains 2806 large aerial images with the size ranges from 800×800 to 4000×4000 and 188, 282 instances among 15 common categories: Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Small vehicle (SV), Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HA), Swimming pool (SP), and Helicopter (HC). **DOTA-v1.5** is released for DOAI Challenge 2019[3] with a new category, Container Crane (CC) and more

[3] https://captain-whu.github.io/DOAI2019

extremely small instances (less than 10 pixels). DOTA-v1.5 contains 402, 089 instances. Compared with DOTA-v1.0, DOTA-v1.5 is more challenging but stable during training.

Following the settings in previous methods [7, 10], we use both training and validation sets for training and the test set for testing. We crop the original images into 1024×1024 patches with a stride of 824. Random horizontal flipping is adopted to avoid over-fitting during training, and no other tricks are utilized. For fair comparisons with other methods, we prepare multi-scale data at three scales $\{0.5, 1.0, 1.5\}$, and random rotation for training and testing.

HRSC2016 [21] is a challenging ship detection dataset with OBB annotations, which contains 1061 aerial images with the size ranges from 300×300 to 1500×900. It includes 436, 181 and 444 images in the training, validation and test set, respectively. We use both training and validation sets for training and the test set for testing. All images are resized to (800, 512) without changing the aspect ratio. Random horizontal flipping is applied during training.

5.2. Implementation Details

ImageNet pretrain. For the original ResNet [12], we directly use the ImageNet pretrained models from Pytorch [25]. For ReResNet, we implement it based on the `mmclassification`[4]. We train ReResNet on the ImageNet-1K with an initial learning rate of 0.1. All models are trained for 100 epochs and the learning rate is divided by 10 at $\{30, 60, 90\}$ epochs. The batch size is set to 256.

Fine-tuning on detection. We adopt ResNet [12] with FPN [17] as the backbone of the baseline method. ReResNet with ReFPN is adopted as the backbone of our proposed ReDet. For RPN, we set 15 anchors per location of each pyramid level. For R-CNN, we sample 512 RoIs with a 1:3 positive to negative ratio for training. For testing, we adopt 10000 RoIs (2000 for each pyramid level) before NMS and 2000 RoIs after NMS. We adopt the same training schedules as `mmdetection` [3]. SGD optimizer is adopted with an initial learning rate of 0.01, and the learning rate is divided by 10 at each decay step. The momentum and weight decay are 0.9 and 0.0001, respectively. We train all models in 12 epochs for DOTA and 36 epochs for HRSC2016. We use 4 V100 GPUs with a total batch size of 8 for training and a single V100 GPU for inference.

5.3. Ablation Studies

In this section, we conduct a series of ablation experiments on DOTA-v1.5 test set to evaluate the effectiveness of our proposed method. Note that we use the original ResNet+FPN and RRoI Align as the backbone and RoI warping method for the baseline method, respectively.

Rotation-equivariant backbone. We evaluate the effectiveness of rotation-equivariant backbone with ReResNet50+ReFPN under different settings. As shown in Tab. 1, compared to ResNet50, ReResNet50 achieves lower clas-

[4] https://github.com/open-mmlab/mmclassification

backbone	group	cls. (%)	det. (%)	size (Mb)
R50-FPN	-	**76.55**	65.03	103
ReR50-ReFPN	C_4	72.81	65.43	24
ReR50-ReFPN	C_8	71.20	**66.86**	12
ReR50-ReFPN	C_{16}	61.60	64.36	**6**

Table 1. **Performance comparisons of the rotation-equivariant backbone on classification (cls.) and detection (det.)**. group indicates the rotation group that the backbone is equivariant to. We report the top-1 accuracy on ILSVRC 2012 without FPN and the detection performance on DOTA-v1.5 test set in terms of mAP. The model size only includes the size of the backbone.

method	backbone	mAP (%)	size (Mb)
FR-O	R50-FPN	62.00	158
	ReR50-ReFPN	**62.36**	**68**
RetinaNet-O	R50-FPN	58.74	140
	ReR50-ReFPN	**59.64**	**34**

Table 2. **The performance of rotation-equivariant backbone on other detectors**. Faster R-CNN OBB (FR-O) and RetinaNet OBB (RetinaNet-O) are our re-implemented version for OBBs.

method	#interpolate	mAP (%)
RRoI Align	-	65.99
RRoI Align+*MP.*	-	64.60 (-1.39)
RiRoI Align	1	66.44 (+0.45)
RiRoI Align	2	**66.86 (+0.87)**
RiRoI Align	4	66.32 (+0.33)

Table 3. **Comparisons of our RiRoI Align with RRoI Align**. #interpolate indicates the number of orientation channels used for interpolation (same as l in Sec. 4.2). For an RRoI with the orientation θ, we use its nearest $\{1, 2, 4\}$ orientation channels to interpolate its features. *MP.* is short for MaxPool. ReR50+ReFPN is adopted as the backbone.

method	rot.	schd.	mAP (%)	training (h)
ReDet	×	1x	62.62	**8**
baseline	✓	1x	64.07	11
ReDet*	×	1x	66.66	13
baseline	✓	2x	**67.34**	22

Table 4. **Comparison with rotation augmentation**. We compare the performance of the baseline method with rotation (rot.) augmentation and ReDet without rotation augmentation. ReDet* preserves a similar amount of parameters with the baseline. We report the mAP with R18 (for baseline) and ReR18 (for ReDet) backbone under the cyclic group C_8. For fair comparison, we randomly select rotation angles from $\{0, 45, 90, \cdots, 315\}$.

method	DOTA-v1.0			HRSC2016		
	AP50	AP75	mAP	AP50	AP75	mAP
baseline	75.62	48.37	46.13	90.18	80.48	68.17
ReDet	76.25	50.86	**47.11(+0.98)**	90.46	89.46	**70.41(+2.24)**

Table 5. **Performance of the proposed ReDet on other datasets**. We report the performance on DOTA-v1.0 and HRSC2016 in COCO style. We use ReR50+ReFPN (*resp.* R50+FPN) as the backbone of ReDet (*resp.* baseline).

method	backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
single-scale:																	
FR-O [35]	R101	79.42	77.13	17.70	64.05	35.30	38.02	37.16	89.41	69.64	59.28	50.30	52.91	47.89	47.40	46.30	54.13
ICN [1]	R101-FPN	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.20	53.64	62.90	67.02	64.17	50.23	68.16
CADNet [42]	R101-FPN	87.80	82.40	49.40	73.50	71.10	63.50	76.60	90.90	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
DRN [24]	H-104	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70
CenterMap [30]	R50-FPN	88.88	81.24	53.15	60.65	78.62	66.55	78.10	88.83	77.80	83.61	49.36	66.19	72.10	72.36	58.70	71.74
SCRDet [40]	R101-FPN	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
R^3Det [37]	R152-FPN	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74
S^2A-Net [10]	R50-FPN	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
ReDet (Ours)	ReR50-ReFPN	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25
multi-scale:																	
RoI Trans.* [7]	R101-FPN	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
O^2-DNet* [31]	H104	89.30	83.30	50.10	72.10	71.10	75.60	78.70	90.90	79.90	82.90	60.20	60.00	64.60	68.90	65.70	72.80
DRN* [24]	H104	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
Gliding Vertex* [36]	R101-FPN	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
BBAVectors* [41]	R101	88.63	84.06	52.13	69.56	78.26	80.40	88.06	90.87	87.23	86.39	56.11	65.62	67.10	72.08	63.96	75.36
CenterMap* [30]	R101-FPN	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
CSL* [38]	R152-FPN	90.25	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
SCRDet++* [39]	R152-FPN	88.68	85.22	54.70	73.71	71.92	84.14	79.39	90.82	87.04	86.02	67.90	60.86	74.52	70.76	72.66	76.56
S^2A-Net* [10]	R50-FPN	88.89	83.60	57.74	81.95	79.94	83.19	89.11	90.78	84.87	87.81	70.30	68.25	78.30	77.01	69.58	79.42
ReDet* (Ours)	ReR50-ReFPN	88.81	82.48	60.83	80.82	78.34	86.06	88.31	90.87	88.77	87.03	68.65	66.90	79.26	79.71	74.67	80.10

Table 6. **Comparisons with state-of-the-art methods on DOTA-v1.0 OBB Task**. H-104 means Hourglass 104. * indicates multi-scale training and testing. The results with red and blue colors indicate the best and second-best results of each column, respectively.

method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC	mAP
OBB results:																	
RetinaNet-O [18]	71.43	77.64	42.12	64.65	44.53	56.79	73.31	90.84	76.02	59.96	46.95	69.24	59.65	64.52	48.06	0.83	59.16
FR-O [27]	71.89	74.47	44.45	59.87	51.28	68.98	79.37	90.78	77.38	67.50	47.75	69.72	61.22	65.28	60.47	1.54	62.00
Mask R-CNN [11]	76.84	73.51	49.90	57.80	51.31	71.34	79.75	90.46	74.21	66.07	46.21	70.61	63.07	64.46	57.81	9.42	62.67
HTC [2]	77.80	73.67	51.40	63.99	51.54	73.31	80.31	90.48	75.12	67.34	48.51	70.63	64.84	64.48	55.87	5.15	63.40
OWSR* [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	74.90
ReDet (Ours)	79.20	82.81	51.92	71.41	52.38	75.73	80.92	90.83	75.81	68.64	49.29	72.03	73.36	70.55	63.33	11.53	66.86
ReDet* (Ours)	88.51	86.45	61.23	81.20	67.60	83.65	90.00	90.86	84.30	75.33	71.49	72.06	78.32	74.73	76.10	46.98	**76.80**
HBB results:																	
RetinaNet-O [18]	71.66	77.22	48.71	65.16	49.48	69.64	79.21	90.84	77.21	61.03	47.30	68.69	67.22	74.48	46.16	5.78	62.49
FR-O [27]	71.91	71.60	50.58	61.95	51.99	71.05	80.16	90.78	77.16	67.66	47.93	69.35	69.51	74.40	60.33	5.17	63.85
HTC [2]	78.41	74.41	53.41	63.17	52.45	63.56	79.89	90.34	75.17	67.64	48.44	69.94	72.13	74.02	56.42	12.14	64.47
Mask R-CNN [11]	78.36	77.41	53.36	56.94	52.17	63.60	79.74	90.31	74.28	66.41	45.49	71.32	70.77	73.87	61.49	17.11	64.54
OWSR* [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	77.90
ReDet (Ours)	79.51	82.63	53.81	69.82	52.76	75.64	87.82	90.83	75.81	68.78	49.11	71.65	75.57	75.17	58.29	15.36	67.66
ReDet* (Ours)	88.68	86.57	61.93	81.20	73.71	83.59	90.06	90.86	84.30	75.56	71.55	71.86	83.93	80.38	75.62	49.55	**78.08**

Table 7. **Performance comparisons on DOTA-v1.5 test set**. Note the results of Faster R-CNN OBB (FR-O) [27], RetinaNet OBB (RetinaNet-O) [18], Mask R-CNN [11] and Hybrid Task Cascade (HTC) [2] are our re-implemented version for DOTA. OWSR [15] is a method from DOAI 2019, and we report its single model performance for fair comparisons. The HBB results of our method are converted from OBB results by calculating the axis-aligned bounding boxes. * means multi-scale training and testing.

sification accuracy due to the reduction of parameters, but it obtains higher detection mAP. We find the backbone under the cyclic group C_8 achieves better accuracy-parameter trade-off. ReResNet50+ReFPN under C_8 gains **1.83** detection mAP improvements with only **1/8** parameters (103 Mb *vs.* 12 Mb). Besides, we also extend ReResNet+ReFPN to other methods in Tab. 2. Both Faster R-CNN OBB and RetinaNet OBB with ReResNet50+ReFPN outperform its counterpart which further demonstrates the effectiveness of rotation-equivariant backbones.

Effectiveness of RiRoI Align. As shown in Tab. 3, compared with RRoI Align, RiRoI Align shows significant improvements due to its orientation alignment mechanism. While RRoI Align+MaxPool leads to a significant drop in mAP, indicating that the orientation pooling is undesirable in oriented object detection. RiRoI Align with a $l = 2$ interpolation achieves the highest **66.86** mAP and **0.87** mAP improvements than RRoI Align. Besides, we find RiRoI Align with a $l = 4$ interpolation only gains **0.33** mAP. The reason may be that too many interpolations hurt the equivariant property and inner relation between orientations.

Comparison with rotation augmentation. From another perspective, our method can be viewed as a special in-network rotation augmentation, which learns from one orientation and can be applied to multiple orientations. In contrast, rotation augmentation enhances the network by generating samples with more orientations and usually requires more time to converge. As shown in Tab. 4, although our method does not exceed the rotation augmented baseline under 1x schedule, our ReDet*, which preserves the similar amount of parameters, shows **2.59** mAP improvements with only **18%** extra training time. Moreover, the 2x baseline with rotation augmentation is **0.68** higher than our ReDet*, but it takes **twice** the training time.

Performance on other datasets. To prove the generalization of our proposed method, we also evaluate the per-

method	RC2 [19]	RRPN [22]	R^2PN [43]	RRD [16]	RoI Trans. [7]	Gliding Vertex [36]
mAP	75.7	79.08	79.6	84.3	86.2	88.2
method	R^3Det [37]	DRN [24]	CenterMap [30]	CSL [38]	S^2A-Net [10]	ReDet (Ours)
mAP	89.26	92.7*	92.8*	89.62	90.17 / 95.01*	**90.46 / 97.63***

Table 8. **Comparisons of state-of-the-art methods on HRSC2016**. * indicates that the result is evaluated under VOC2012 metrics, while other methods are all evaluated under VOC2007 metrics. We report both results for fair comparisons.

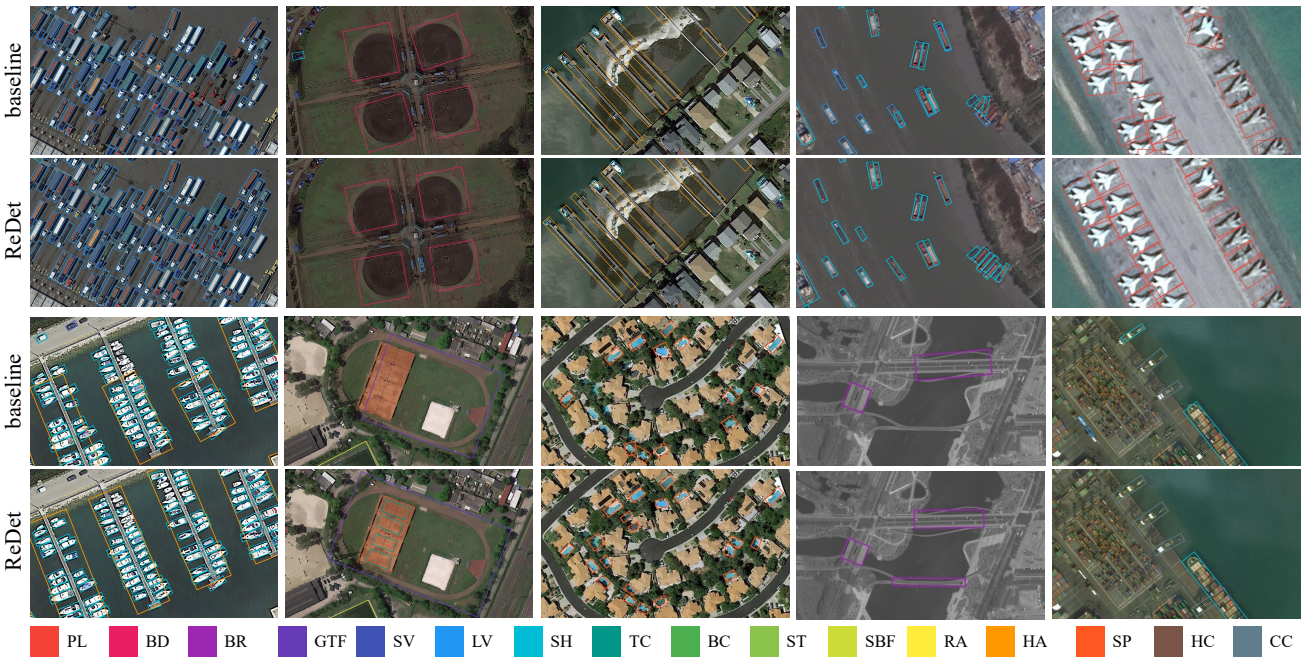

Figure 4. **Qualitative comparisons** between the proposed ReDet and the baseline method on DOTA-v1.5.

formance of ReDet on DOTA-v1.0 and HRSC2016. As is shown in Tab. 5, compared with the baseline, ReDet achieves better performance on both datasets. Moreover, ReDet has significant improvements in **AP75** and **mAP**, which demonstrates its accurate localization capabilities.

5.4. Comparisons with the State-of-the-Art

Results on DOTA-v1.0. As shown in Tab. 6, we compare our ReDet with other state-of-the-art methods on DOTA-v1.0 OBB Task. Without bells and whistles, our single-scale model achieves **76.25** mAP, outperforming all single-scale models and most multi-scale models. With limited data augmentation (*i.e.*, multi-scale data and random rotation), our method achieves state-of-the-art **80.10** mAP in the whole dataset, and obtains the best or second-best results among **12/15** categories.

Results on DOTA-v1.5. Compared with DOTA-v1.0, DOTA-v1.5 contains many extremely small instances, which increases the difficulty of object detection. We report both OBB and HBB results on DOTA-v1.5 test set in Tab. 7. With single-scale data, our method achieves **66.86** OBB mAP and **67.66** HBB mAP, outperforming RetinaNet OBB, Faster R-CNN OBB, Mask R-CNN [11] and HTC [2] by a large margin. Especially for the categories with small instances (*e.g.*, HA, SP, CC) and large scale variations (*e.g.*, PL, BD), our method performs better. Besides, as shown in Fig. 2, our ReDet achieves better parameter *vs.* accuracy trade-off, which further demonstrates its efficiency. Compared to previous best results by OWSR [15], our multi-scale model achieves state-of-the-art performance, about **76.80** OBB mAP and **78.08** HBB mAP. Qualitative comparisons between our ReDet and the baseline method are visualized in Fig. 4.

Results on HRSC2016. The HRSC2016 contains a lot of thin and long ship instances with arbitrary orientation. We compare our ReDet with other state-of-the-art methods in Tab. 8. Our method achieves the state-of-the-art performance, *i.e.*, with mAP of **90.46** and **97.63** under the VOC2007 and VOC2012 metrics, respectively.

6. Conclusions

This paper presents a Rotation-equivariant Detector for aerial object detection, which consists of two parts: the rotation-equivariant backbone and the RiRoI Align. The former produces rotation-equivariant features, while the latter extracts rotation-invariant features from rotation-equivariant features. Extensive experiments on DOTA and HRSC2016 demonstrate the effectiveness of our method.

References

[1] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *ACCV*, pages 150–165, 2018. 2, 7

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, pages 4974–4983, 2019. 2, 7, 8

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[4] G. Cheng, P. Zhou, and J. Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2016. 3

[5] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, pages 2990–2999, 2016. 2, 3, 4, 5

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 3

[7] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *CVPR*, pages 2849–2858, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[8] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 2, 3

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 2

[10] J. Han, J. Ding, J. Li, and G. S. Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–11, 2021. 1, 6, 7, 8

[11] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017. 2, 7, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 6

[13] Emiel Hoogeboom, Jorn WT Peters, Taco S Cohen, and Max Welling. Hexaconv. In *ICLR*, 2018. 2, 3

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015. 3

[15] Chengzheng Li, Chunyan Xu, Zhen Cui, Dan Wang, Zequn Jie, Tong Zhang, and Jian Yang. Learning object-wise semantic representation for detection in remote sensing imagery. In *CVPR workshops*, pages 20–27, 2019. 7, 8

[16] Minghui Liao, Zhen Zhu, Baoguang Shi, Guisong Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *CVPR*, pages 5909–5918, 2018. 8

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5, 6

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 2, 7

[19] Lei Liu, Zongxu Pan, and Bin Lei. Learning a rotation invariant detector with rotatable bounding box. *arXiv preprint arXiv:1711.09405*, 2017. 8

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016. 2

[21] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *ICPRAM*, pages 324–331, 2017. 1, 2, 6

[22] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. on Multimedia*, 2018. 1, 2, 3, 4, 8

[23] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *ICCV*, pages 5048–5057, 2017. 3

[24] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. In *CVPR*, June 2020. 2, 3, 7, 8

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 6

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. on PAMI*, pages 1137–1149, 2017. 2, 7

[28] Yun Ren, Changren Zhu, and Shunping Xiao. Deformable faster r-cnn with aggregating multi-layer features for partially occluded object detection in optical remote sensing images. *Remote Sensing*, 10(9):1470, 2018. 3

[29] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *CVPR*, pages 4168–4176, 2016. 3

[30] Jinwang Wang, Wen Yang, Heng-Chao Li, Haijian Zhang, and Gui-Song Xia. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 3, 7, 8

[31] Haoran Wei, Yue Zhang, Zhonghan Chang, Hao Li, Hongqi Wang, and Xian Sun. Oriented objects as pairs of middle lines. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:268–279, 2020. 3, 7

[32] Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In *NeurIPS*, pages 14334–14345, 2019. 3, 5

[33] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *CVPR*, pages 849–858, 2018. 2, 3, 4, 5

[34] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukham-

betov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *CVPR*, pages 5028–5037, 2017. 3

[35] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, pages 3974–3983, 2018. 1, 2, 5, 7

[36] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Guisong Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. on PAMI*, 2020. 3, 7, 8

[37] Xue Yang, Qingqing Liu, Junchi Yan, Ang Li, Zhiqiang Zhang, and Gang Yu. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv preprint arXiv:1908.05612*, 2019. 1, 7, 8

[38] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *ECCV*, 2020. 1, 2, 3, 7, 8

[39] Xue Yang, Junchi Yan, Xiaokang Yang, Jin Tang, Wenlong Liao, and Tao He. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *arXiv preprint arXiv:2004.13316*, 2020. 7

[40] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *ICCV*, pages 8231–8240, 2019. 1, 2, 7

[41] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. Oriented object detection in aerial images with box boundary-aware vectors. *arXiv preprint arXiv:2008.07043*, 2020. 3, 7

[42] Gongjie Zhang, Shijian Lu, and Wei Zhang. Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, PP:1–10, 2019. 2, 7

[43] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geoscience and Remote Sensing Letters*, (99):1–5, 2018. 2, 8

[44] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 3

[45] Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *CVPR*, pages 4961–4970, 2017. 3, 5